Review of Disordered Voice Enhancement

Engineering Approaches

Martin Hagmüller
hagmueller@tugraz.at

Signal Processing and Speech Communication Laboratory
Graz University of Technology – Austria

http://www.spsc.tugraz.at
Introduction
Healthy Vocal Cords
Disordered Voices - Perceptual Effects

- **Hoarseness:**
  Perceptual cue of voice disorder
Disordered Voices - Perceptual Effects

- **Hoarseness:**
  Perceptual cue of voice disorder

- **Roughness:**
  Perceptual cue of irregularity of the vocal fold oscillation
  - Sound example: 🎧
  
  can be caused by:
  - Surgically removed larynx (→ no vocal folds)
  - Vocal fold nodules
  - ...

**Techniques Used**
- EL Model
- Multipath Separation
- F0 for EL
- Excitation Signal
Disordered Voices - Perceptual Effects

- **Hoarseness:**
  Perceptual cue of voice disorder

- **Roughness:**
  Perceptual cue of irregularity of the vocal fold oscillation
  - Sound example: 
  can be caused by:
  - Surgically removed larynx (→ no vocal folds)
  - Vocal fold nodules
  - ...

- **Breathiness:**
  Excessive amount of turbulent airflow in voiced speech
  - Sound example: 
  can be caused by:
  - Vocal fold immobility
  - Vocal fold nodules
  - ...

- Hoarseness:
- Roughness:
- Breathiness:
Disordered Voices - Causes

- **Organic voice disorder:**
  - Physical abnormality of larynx
    (e.g.: polyp, nodule, cancer)
    Video: [Video]
  - Neurological problem
    (e.g. Vocal fold paralysis, Parkinson disease)
    Video: [Video]
Disordered Voices - Causes

- Organic voice disorder:
  - Physical abnormality of larynx (e.g.: polyp, nodule, cancer)
    Video: [Video]
  - Neurological problem (e.g. Vocal fold paralysis, Parkinson disease)
    Video: [Video]

- Functional voice disorder:
  - Misuse of vocal organ (e.g. overuse)
  - Psychological causes (e.g. juvenile voice)
Disordered Voices - Causes

- **Organic voice disorder:**
  - Physical abnormality of larynx
    (e.g.: polyp, nodule, cancer)
  - Video: [Play]
  - Neurological problem
    (e.g. Vocal fold paralysis, Parkinson disease)
  - Video: [Play]

- **Functional voice disorder:**
  - Misuse of vocal organ (e.g. overuse)
  - Psychological causes (e.g. juvenile voice)

- **Substitution Voice:**
  - No larynx
Substitution Voices

- Esophageal speech (ES)
Substitution Voices

- Esophageal speech (ES)
- Tracheo-Esophageal speech (TE)

Sound example: 🎧
Substitution Voices

- Esophageal speech (ES)
- Tracheo-Esophageal speech (TE)
- Electrolarynx speech (EL)
Issues

- Voicing Source
  - Unnatural Voicing (EL, TE, ES)
  - Low harmonics-to-noise ratio (HNR) due to irregular voicing source
  - Mechanic sound (EL)
  - No or limited pitch variation (EL)
  - Low Pitch (ES, TE) - Particularly Female!
Issues

- Voicing Source
  - Unnatural Voicing (EL, TE, ES)
  - Low harmonics-to-noise ratio (HNR) due to irregular voicing source
  - Mechanic sound (EL)
  - No or limited pitch variation (EL)
  - Low Pitch (ES, TE) - Particularly Female!

- Additional Noise
  - Injection Noise (ES)
  - Stoma Noise (ES, TE, ES)
  - Electro-larynx Noise (EL)
Issues

- **Voicing Source**
  - Unnatural Voicing (EL, TE, ES)
  - Low harmonics-to-noise ratio (HNR) due to irregular voicing source
  - Mechanic sound (EL)
  - No or limited pitch variation (EL)
  - Low Pitch (ES, TE) - Particularly Female!

- **Additional Noise**
  - Injection Noise (ES)
  - Stoma Noise (ES, TE, ES)
  - Electro-larynx Noise (EL)

- **Modified anatomy (EL, ES, TE)**
  - Formant shifted upwards
  - Articulation problems due to muscular support for tongue
Issues

- Voicing Source
  - Unnatural Voicing (EL, TE, ES)
  - Low harmonics-to-noise ratio (HNR) due to irregular voicing source
  - Mechanic sound (EL)
  - No or limited pitch variation (EL)
  - Low Pitch (ES, TE) - Particularly Female!

- Additional Noise
  - Injection Noise (ES)
  - Stoma Noise (ES, TE, ES)
  - Electro-larynx Noise (EL)

- Modified anatomy (EL, ES, TE)
  - Formant shifted upwards
  - Articulation problems due to muscular support for tongue

- Hand is needed for talking (EL, TE)
Implications of alaryngeal voice

- Severe handicap for social interaction
  - Weak voice
  - Unnatural voice
  - Loss of voice, a traumatic experience
Implications of alaryngeal voice

- Severe handicap for social interaction
  - Weak voice
  - Unnatural voice
  - Loss of voice, a traumatic experience

- Problems for digital speech processing
  - Speech compression used for digital telephony // (e.g. GSM - Mobile Phone)
  - Speech recognition
Implications of alaryngeal voice

- Severe handicap for social interaction
  - Weak voice
  - Unnatural voice
  - Loss of voice, a traumatic experience

- Problems for digital speech processing
  - Speech compression used for digital telephony // (e.g. GSM - Mobile Phone)
  - Speech recognition

- Applications
  - Add on device for telephone communication
  - Device for acoustically difficult situations (loud environment)
Techniques Used

- Framework
- Voicing substitution-enhancement
- Voice conversion
- Noise Reduction
- E-Larynx Efficiency Improvement
- Enhanced Voice Production System with E-Larynx
- Laryngeal Speech

EL Model

Multipath Separation

F0 for EL

Excitation Signal

Conclusion

Appendix
Framework

Analysis - Modification - Resynthesis

Speech → Analysis/Transformation → Modification → Resynthesis Inverse Transformation → Enhanced Speech
Introduction

Techniques Used
- Framework
- Voicing substitution-enhancement
- Voice conversion
- Noise Reduction
- E-Larynx Efficiency Improvement
- Enhanced Voice Production System with E-Larynx
- Laryngeal Speech

EL Model
- Multipath Separation
- F0 for EL
- Excitation Signal

Conclusion

Appendix

Framework

Analysis - Modification - Resynthesis

Speech → Analysis/Transformation → Modification → Resynthesis Inverse Transformation → Enhanced Speech

- Linear Prediction Analysis
  [Qi et al., 1995, Ali and Jebara, 2006, Pozo and Young, 2006]

- Modification of Parameters

- Linear Prediction Analysis

Pitch period

Impulse train generator

Voiced/unvoiced switch

Random noise generator

Vocal tract parameters

Speech output

Vocal tract model

gain
Framework

Analysis - Modification - Resynthesis

- Linear Prediction Analysis
  
  [Qi et al., 1995, Ali and Jebara, 2006, Pozo and Young, 2006]

- Modification of Parameters

- Fourier Transform - Frequency Domain [Liu et al., 2006]
Framework

Analysis - Modification - Resynthesis

- Linear Prediction Analysis
  [Qi et al., 1995, Ali and Jebara, 2006, Pozo and Young, 2006]
- Modification of Parameters
- Fourier Transform - Frequency Domain [Liu et al., 2006]
- Voice Pulse Modeling [Loscos and Bonada, 2006]
Voicing substitution-enhancement

- Excitation source synthesis
Voicing substitution-enhancement

- Excitation source synthesis
  - Liljencrantz-Fant (LF) model
    [Qi et al., 1995, Bi and Qi, 1997, Ali and Jebara, 2006]
    Glottal flow determined by 4 independent parameters
Voicing substitution-enhancement

- Excitation source synthesis
  - Liljencrants-Fant (LF) model
    - Glottal flow determined by 4 independent parameters
      - [Qi et al., 1995, Bi and Qi, 1997, Ali and Jebara, 2006]
  - Healthy excitation source sample
    - [Matsui and Hara, 1999, Houston et al., 1999]
Voicing substitution-enhancement

- Excitation source synthesis
  - Liljencrans-Fant (LF) model
    - Glottal flow determined by 4 independent parameters
      
      ![Graph](image)

    - Healthy excitation source sample
      - Matsui and Hara, 1999, Houston et al., 1999

- Artificial pitch from energy envelope
  - Loscos and Bonada, 2006
Voice conversion

Formant frequencies are moved to correct position

Learning Algorithm

[Li and Qi, 1997, Nakamura et al., 2006, Aguilar-Torres et al., 2006]
Voice conversion

Formant frequencies are moved to correct position

- **Learning Algorithm**

  - Bi and Qi, 1997, Nakamura et al., 2006, Aguilar-Torres et al., 2006

  ![Diagram showing voice conversion process](image)

  - **Formant modification by some calculation**

    - Ali and Jebara, 2006, Loscos and Bonada, 2006, Pozo and Young, 2006

    - Formant shift
    - Bandwidth changed
    - High frequency de-emphasis

![Graph showing frequency analysis](image)
Noise Reduction

- **Spectral Smoothing** [Matsui and Hara, 1999, Pozo and Young, 2006] Linear prediction parameter smoothing

![Graph showing spectral smoothing](image)
Noise Reduction

- **Spectral Smoothing** [Matsui and Hara, 1999; Pozo and Young, 2006] Linear prediction parameter smoothing
- **Injection noise rejection** - [Javkin et al., 1997]
Noise Reduction

- Spectral Smoothing \[\text{[Matsui and Hara, 1999; Pozo and Young, 2006]}\] Linear prediction parameter smoothing

- Injection noise rejection - \[\text{[Javkin et al., 1997]}\]

- Spectral subtraction - \[\text{[Cole et al., 1997]}\] using auditory masking - \[\text{[Liu et al., 2006]}\]
Noise Reduction

- **Spectral Smoothing** [Matsui and Hara, 1999; Pozo and Young, 2006] Linear prediction parameter smoothing
- **Injection noise rejection** - [Javkin et al., 1997]
- **Spectral subtraction** - [Cole et al., 1997] using auditory masking - [Liu et al., 2006]
- **Adaptive Filtering** - [Espy-Wilson et al., 1998, Niu et al., 2003]
E-Larynx Efficiency Improvement

- E-Larynx with linear transducer - [Houston et al., 1999]
E-Larynx Efficiency Improvement

- E-Larynx with linear transducer - [Houston et al., 1999]
- Impedance matching
E-Larynx Efficiency Improvement

- E-Larynx with linear transducer - [Houston et al., 1999]

- Impedance matching

- Incorporating the neck tissue frequency response - [Norton and Bernstein, 1993, Meltzer et al., 2003]
Enhanced Voice Production System with E-Larynx

Non Audible Murmur Microphone - [Nakamura et al., 2006]

- NAM picks up structure borne sound
Enhanced Voice Production System with E-Larynx

- Non Audible Murmur Microphone - [Nakamura et al., 2006]
  - NAM picks up structure borne sound
- Electromyographic E-Larynx Control - [Goldstein et al., 2004]
Enhanced Voice Production System with E-Larynx

- Non Audible Murmur Microphone - [Nakamura et al., 2006]
  - NAM picks up structure borne sound
- Electromyographic E-Larynx Control - [Goldstein et al., 2004]
Hoarseness Reduction for Laryngeal Speech

- **Singular Value Decomposition** [Manfredi et al., 2006]
Hoarseness Reduction for Laryngeal Speech

- **Singular Value Decomposition** [Manfredi et al., 2006]
- **Nonlinear Corrections of Vocal Disorder**
  
  [Matassini and Manfredi, 2001; Hegger et al., 2001] Like averaging over signal of several copies of a long playing record. Noise due to scratches on the disc will be reduced.
Hoarseness Reduction for Laryngeal Speech

- **Singular Value Decomposition** [Manfredi et al., 2006]
- **Nonlinear Corrections of Vocal Disorder**
  - [Matassini and Manfredi, 2001], [Hegger et al., 2001] Like averaging over signal of several copies of a long playing record. Noise due to scratches on the disc will be reduced
- **Periodicity Enhancement and Pitch Correction**
  - [Hagmüller and Kubin, 2004]
  
  **Original:**
  
  **Modified:**
Electro-Larynx Speech Production Model
EL Speech Production

Speech signal path, $h_{nv}(t)$

Direct path, $h_d(t)$
EL Speech Production Model

Random Sound Source: $u(t) \rightarrow h_v(\tau, t) \rightarrow s_u(t)$

Shaker: $e(t) \rightarrow h_{nv}(\tau, t) \rightarrow s_v(t)$

Impulse Generator: $h_d(t)$

Environmental noise: $n(t)$

$y(t) = s_u(t) + s_v(t) + h_d(t) + n(t)$

EL Speech & Noise
EL Speech Spectrogram

- Constant $F_0$
- Strong harmonics
- Example:
Multipath Separation
Introduction

- Suppress directly radiated EL noise
- E.g., adaptive filtering \cite{Espy-Wilson1998, Niu2003}
- But: Independence of signal & noise cannot be assumed
Introduction

- Suppress directly radiated EL noise
- E.g., adaptive filtering [Espy-Wilson et al., 1998, Niu et al., 2003]
- But: Independence of signal & noise cannot be assumed
Multipath Separation

- Time varying speech path $h_{nv}(\tau, t)$
- Time invariant noise path $h_d(t)$

![Diagram showing the process of multipath separation]
Multipath Separation

- Time varying speech path  \( h_{nv}(\tau, t) \)
- Time invariant noise path  \( h_d(t) \)

→ Modulation Frequency Filtering

\[
\begin{align*}
\text{Random Sound Source} & \rightarrow h_v(\tau, t) \rightarrow s_v(t) \\
\text{Shaker} & \rightarrow h_{nv}(\tau, t) \rightarrow s_v(t) \\
\text{Impulse Generator} & \rightarrow h_d(t) \rightarrow s_d(t) \\
\text{Environment noise} & \rightarrow n(t) \\
\end{align*}
\]

\[
\begin{align*}
\text{Modulator} & \rightarrow m_1[n] \\
\text{Carrier} & \rightarrow c_1[n] \\
\text{Modulation Filter} & \rightarrow y_1[n] \\
\text{Reconstruction filter bank} & \rightarrow y[n] \\
\end{align*}
\]
Objective Evaluation

Spectrogram of EL speech. ‘zwei drei’.

- SNR improvement of 15-20 dB
Results

- Naive Listeners: Preference of modified EL speech
- Speech Therapists: Strong preference of original EL speech
- Laryngeal vs. Alaryngeal Speakers: No significant difference

<table>
<thead>
<tr>
<th>Original:</th>
<th>Multipath Separation:</th>
</tr>
</thead>
</table>
Prosody for Alaryngeal Speech
Previous work

- Pitch Control for E-Larynx
  - Button to switch F0 to place accent
  - Pressure sensitive button to control F0
    → Hard to control

- Artificial Pitch Contour from energy envelope
  - Tonal languages: No intelligibility loss in whispered speech
    [Meyer-Eppler, 1956]
  - Whisper: Correlation of Formants and Perceived Pitch
    [Higashikawa and Minifie, 1999]
  - Alaryngeal speakers: Ability to convey prosodic information
    [van Rossum et al., 2002]
Formants in Singing

Introduction

Techniques Used

EL Model

Multipath Separation

F0 for EL

■ Previous work
■ Formants in Singing
■ Applications
■ Approach
■ F0 Generation
■ Evaluation
■ Results

Excitation Signal

Conclusion

Appendix

April 2011

Voice Enhancement - p. 28/49
Applications

$F_0$ control of EL

$F_0$ modification of EL speech
Approach

- Declination: Long-term downward trend based on EL activity
- Local $F_0$ contour from formants
F0 Generation

Fujisaki Model

1. EL is
   \[ F_p(t) = \begin{cases} 
   Ft(1 - e^{-\tau_{\text{rise}} t}) & \text{off} \\
   Ft e^{-\tau_{\text{fall}} t} & \text{on} 
   \end{cases} \]

2. \[ F_a(t) = (F_2(t) - \text{mean}(F_2(t))) \cdot \beta + \gamma, \]

3. \[ F_0(t) = F_a(t) + F_p(t) \]
Evaluation

- Pitch modification done with PSOLA

- Speakers:
  - 9 healthy subjects
    - Laryngeal
    - Whisper
    - EL Speech

- Subjective Evaluation
  - Linguistic Function
    - Emphasis Position
    - Sentence Mode
  - Preference of $F_0$ contour
Results

Original:  

Artificial $F_0$:  

\textit{Nordwind und Sonne} ...

- Linguistic Function: No enhancement
- $F_0$ Contour: Significant preference of generated $F_0$ contour
Excitation Signal Models
Approach

- Better excitation signal for EL
  - Linear transducer vs. non-linear transducer

- Voice production models
  - Physical models
  - Waveform models
Models under Investigation

- **Physical models**
  - Delayed one-mass model ($D_1$)
    Mucosal wave modelled with a transmission line
  - Delayed one-mass model ($D_2$)
    Flow generation with learning of non-linear mapping from recorded speech data

- **Waveform models**
  - R++ model
    Parametrical model derived form LF-model
  - Model for disordered voices (HGS)
    Model based on non-linear shaping functions
  - Non-linear oscillator-plus-noise model (O+N)
    Autonomous non-linear system learned from recorded signals
  - Van der Pol Oscillator (VdP)
Consideration of Signal Path

- Shaker transfer function
- Neck transfer function

![Diagram showing signal processing components: DAC, EDIROL UA-25, ALESIS RA300, Oral Cavity, Neck Tissue]
Recordings

Recording Setup

- Recording studio
- Omni-directional condenser head-set microphone (AKG HC-577)
  - \( f_s = 44.1 \text{kHz} \)
- Linear transducer: Brüel & Kjær 4810 (shaker)

Procedure

- 2 healthy male speakers
- Oldenburger Satztest (10 sentences)
- 6 excitation signal models + Servox EL device
Listening examples

Servox:  
VdP-model:  
D₁-model:  
D₂-model:  
R++-model:  
HGS-model:  
O+N-model:  

Peter bekommt drei große Blumen
Listening Test

- Qualities assessed:
  - Voice Spectrum/Quality
  - Directly radiated noise
  - Listening effort
  - Overall quality

- Comparison category rating (CCR)
  - 88 randomized sentences
    - 10 sentences
    - 7 excitation signals
    - 2 speakers
    - Null pairs included
    - 980 possible pairs
    - Every listener hears all possible excitation signal combinations

- 20 listeners – mean age: 27 years – 16 male / 4 female

- CMOS results → order of preference

![Comparison Mean Opinion Score]

<table>
<thead>
<tr>
<th>Rating</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>3:</td>
<td>much better</td>
</tr>
<tr>
<td>2:</td>
<td>better</td>
</tr>
<tr>
<td>1:</td>
<td>slightly better</td>
</tr>
<tr>
<td>0:</td>
<td>about the same</td>
</tr>
<tr>
<td>-1:</td>
<td>slightly worse</td>
</tr>
<tr>
<td>-2:</td>
<td>worse</td>
</tr>
<tr>
<td>-3:</td>
<td>much worse</td>
</tr>
</tbody>
</table>

- Second voice token compared to first one is ...

- Qualities assessed:
  - Voice Spectrum/Quality
  - Directly radiated noise
  - Listening effort
  - Overall quality

- Comparison category rating (CCR)
  - 88 randomized sentences
    - 10 sentences
    - 7 excitation signals
    - 2 speakers
    - Null pairs included
    - 980 possible pairs
    - Every listener hears all possible excitation signal combinations

- 20 listeners – mean age: 27 years – 16 male / 4 female

- CMOS results → order of preference
Results 1/2

- Spectrum comparison for different models:
  - D1, D2, O+N, HGS, R++, VdP, Servox

- Noise level comparison for different models:
  - D1, D2, O+N, HGS, R++, VdP, Servox
Results 2/2

- Listening effort
- Overall quality

Comparison of different models:
- D_1
- D_2
- O+N
- HGS
- R++
- VdP
- Servox

Evaluation criteria:
- Listening effort
- Overall quality
Conclusions

- Current EL-Device:
  + Listening effort
  - Voice Quality, Noise
- Potential for improved naturalness
- Models with jitter evaluated better
- Trade-off: intelligibility - naturalness
- Further improvement with more natural $F_0$ contour
Conclusion
Summary

- Improvement of source
- Noise reduction / EL background noise
- Formant correction
- Pitch variability
Discussion

- Open Topics:
  - Bad signal quality
  - Source signal
Discussion

- Open Topics:
  - Bad signal quality
  - Source signal

- Discussion:
  - Results difficult to compare
  - Mainly single effort works
  - Few long term, multipartner projects
  - ...


Appendix
Auditory Model Spectral Subtraction

- Noisy Speech
- Noise Estimation
- Calculate Perc. Weight Filter
- T(ω)
- Perceptually Weighting Coefficients
- Windowing + FFT
- Magnitude
- Phase
- Parametric Spectrum Subtraction
- IFFT Overlap/Add

- Techniques Used
- EL Model
- Multipath Separation
- F0 for EL
- Excitation Signal
- Conclusion
- Appendix
Auditory Model Spectral Subtraction

Flowchart:

1. Noisy Speech
2. Calculate Perc. Weight Filter
3. Perceptually Weighting Coefficients
4. Noise Estimation
5. Noise
6. Windowing + FFT
7. Windowing + FFT
8. Parametric Spectrum Subtraction
9. Magnitude
10. Phase
11. IFFT Overlap/Add

Flowchart steps:

- Calculate Perc. Weight Filter
- Perceptually Weighting Coefficients
- Noise Estimation
- Windowing + FFT
- Parametric Spectrum Subtraction
- Magnitude
- Phase
- IFFT Overlap/Add
Auditory Model Spectral Subtraction

1. Calculate Perc. Weight Filter
2. Perceptually Weighting Coefficients
3. Noise Estimation
4. Windowing + FFT
5. Magnitude
6. Phase
7. Parametric Spectrum Subtraction
8. IFFT Overlap/Add
9. T(ω)
10. Noise
11. α, β

Noisy Speech → Percept. Spectral Subtraction